Prediction of Next Search Query using Association Rule Mining

Rosy Madaan, Manvi Breja, A.K. Sharma, Ashutosh Dixit

Department of Computer Engineering,

YMCA University of Science and Technology, Faridabad

madaan.rosy@gmail.com

Abstract— **The paper presents the method of predicting the next incoming query that the user provides to the search engine. The approach used extracts the information from the log of the previously submitted queries to the search engine, using algorithms for mining association rules. The paper highlights the novel approach of the association rule mining that predicts the next query that the user will provide to the search engine. Using this approach, the search engine keeps the relevant pages in the repository for providing a speedy response to the user and thus increasing the efficiency of the search engine.**

Keywords— **Search Engine, Query Log, Association rule mining, Apriori.**

I. INTRODUCTION

Web has become an essential source of up-to-date information for the users. Thousands of the users surf the internet to achieve their query results through search engine. Search engine is an information retrieval system [1] designed to minimize the time required to find information over the vast Web consisting of hyperlinked documents. It provides a query interface that enables the users to specify criteria about an item of interest and searches the same from locally maintained databases. The criteria are referred to as a *search query*. In the case of text search engines, the search query is typically expressed as a set of words that identify the desired concept that one or more documents may contain. The search queries are posed in a different way by every user. There are four types of search queries- *Informational Queries[3]* that covers a wide topic and gives thousands of relevant answers, *Navigational Queries[3]* in which these queries are in the form of a single website, *Transactional Queries[3]* that refer to a particular action, like shopping or downloading a screen saver, *Connectivity Queries[3]* in which the queries are based on the connectivity of the indexed web graph.

There are various components of Web search engine: *crawler* module which downloads Web pages, page repository in which the downloaded Web pages are temporarily stored in local storage of search engine, indexing module takes the uncompressed page as input and outputs a compressed version of the page. Another module of the search engine is the *query module* that converts a user's natural language query into a language that the search system can understand and consults the various indexes in order to provide a set of relevant pages as answers to the query. The *ranking module* takes the set of relevant pages and ranks them according to some criteria such as popularity score, content score etc and thus presents sorted results to the user.

The act of the search engine is limited to the problem of "Information Overkill". One of the great challenges faced by the search engines is the difficulty in finding the exact need of the user, as the user enters a short and precise query to the interface. If anyhow the search engine could guess the query that can be posed by the user and what the user wants to know, its performance will definitely improve.

This paper has been organized in following sections: Section 2 describes the current research that has been carried out in this area, Section 3 discusses the proposed work, Section 4 describes the architecture that has been used in the proposed work, Section 5 describes the snapshots of the results of experimental evaluation, Section 6 concludes the discussion.

II. RELATED WORK

Fonseca, Golgher, De Moura, and Ziviani in [2] segmented query sessions in search engine query logs into subsessions and then used association rules to extract related queries from those subsessions. Association rules are widely used to develop high-quality recommendation systems in e-commerce applications available in the Web (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994). These applications take user sessions stored in system logs to obtain information about the user behavior to recommend services and products.

In [3], a novel approach to predict the oncoming query for the search engine has been discussed. This approach uses neural networks for the prediction.
In [4], the authors predict users' future queries and URL clicks based on their current access behaviors and global users' query logs. They also explored various features from queries and clicked URLs in the users' current search sessions, select similar intents from query logs, and use them for prediction. Query Log excerpt (RFP 2006 dataset) has been taken as an experimental corpus. Three methods and the back-off models have been presented.

In [5], a method to help a user redefine a query based on past users experience, namely the click-through data as recorded by a search engine has been presented. The method proposed

attempts to recommend better queries rather than related queries. It is effective at identifying query specialization or sub-topics because it takes into account the co-occurrence of documents in individual query sessions.

III. PROPOSED WORK

The paper presents a novel approach to predict the next query that the user may provide to the search engine. The approach discussed uses *association rule mining*. Given an initial input query, the user can see the list of the queries which he may be interested in posing next to the search engine. For the purpose, the system for query prediction has been proposed.

The proposed system of query prediction uses the following steps:

Step-1 Maintaining the Query Log

First, the proposed system maintains a *query log* for the user. *Query log* consists of transactionId, date, time, queryId and name of the queries entered by the user. The transactions in the *Query log* are maintained on the day to day basis.

A *query log* is basically a file containing the interactions that occurred between the users and the system. The content of the search engine's web server log depends on the server and its settings. The log's entries can be simple or complex and presented in different forms. But in general, the log contains the following information:

- Time and date of transaction.
- Name of the query.
- IP address to which the query was sent.
- How the query was sent?
- Possible link that led to that page.
- Information about the browser.
- Transmission results.
- Rank of the clicked result.

The following is an example of an entry in the Query log.

1042078585.991 3713 200.226.211.142 TCP MISS/200 25368 GET http://cluster.igbusca-cluster/query.cgi?query=+origem+da+familia+marques-DIRECT/192.168..12 text/html

Step-2 Applying the Query Mining System on the Query Log

Now, the query logs of different users, their descriptions, interests and the queries they usually pose to the system at a particular time has been maintained. Section IV describes the proposed "Query Mining System" that can be applied on the query log to predict next the incoming queries to the system.

Step-3 Predicting the queries

Since, each user will have different query log, the user login will be required to identify which user has logged in. The system maintains search interface in which the user will provide the initial query to the. And there is another search box; in which the user will provide the minimum support (frequency) count. Whenever user clicks the button of the related queries, the user will get a list of the next predicted queries. In this way the system, helps the particular user logged in, to get the list of the next queries to help to tune the search process. It not only fastens the search but also increases the efficiency of the system, as it will store the next predicted queries in its cache to provide the desired result page to the user in a very less time.

Thus, Query Prediction is one of the major functionality required for the search engine. Thus, the Query prediction system can be used in the search engine's architecture to predict the next query provided to the search engine by the user on the basis of past queries entered by him

There are two key performance indicators for the web search engines:

- Quality of the returned results
- Speed with which the results are returned.

Query prediction aims to increase the speed with which the search engines respond to the users with the returned results i.e. the response time of the search engine will reduce and thus its efficiency will improve.

IV. QUERY MINING SYSTEM

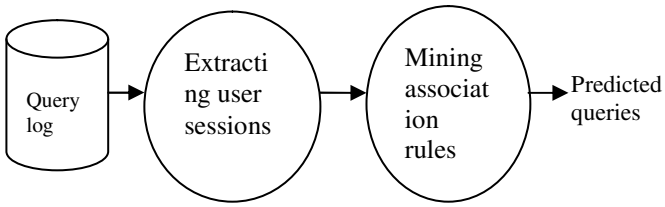

Figure1: Process employed in Query Mining System

The process employed in Query mining system consists of two main functionalities –one for *extracting user sessions* and other for *mining association rules.*

a. Extracting user sessions

User session is basically for how much time the user is active on the server. In our proposed system, the *time* is field is divided into time intervals of say t minutes each and the *user session* is defined by the set of queries each user (identified by the IP address) submits in the time interval. If many queries are submitted within the time interval, then the system restricts the number to a minimum value.

b. Mining Association rules

Association rule mining is one of the data mining techniques which aim to discover the interesting patterns from the large transactional databases. For the system, the technique is applied on the query log that has been maintained.

An association rule is identified as the rule X=>Y where X and Y are set of the items. In our system, X and Y are the queries which are divided into the set of transactions. In the association rule mining, we first do the frequent pattern analysis in which we find the patterns (set of items, subsequences, substructures etc.) that occur frequently in the set.

This can be defined mathematically way as let $I=\{I_1, I_2,.....,I_m\}$ be a set of literals called items. Let T be the database of transactions. Each transaction t can be represented by a binary vector, with t[k]=1 if t bought the item I_k, and t[k]=0 otherwise. Let X is a subset of T. A transaction t satisfies X if for all items I_k in X, t[k]=1.

Similarly, in this way, we can define our problem of finding next predicted queries. Here, the set $Q=\{Q_1, Q_2,.....,Q_m\}$ be a set of queries from the query log database. Let T be the database of set of user sessions. Each session t can be represented by a binary vector, with t[k]=1 if query Q_k fired in session t, and t[k]=0 otherwise. Let X be a subset of T. A transaction t satisfies X if for all queries Q_k in X, t[k]=1.

There are two important parameters that we need to consider.

- **Support**

The rule X=>Y has a support factor of s if s% of the transactions in T that contains X U Y. The problem of mining association rules is to generate all the association rules that have a support greater than a minimum support threshold (*minsup*) set up during the experimental analysis. Thus it shows only those rules in which all queries together appear more number of times than the number set up that is minsup.

- **Confidence**

The rule X=>Y has a confidence factor of c if c% of the transactions in T that contains X also contains Y. This means the rule like Q_1=>Q_2 has a confidence factor c if in c% of the total transactions in database, if Q_1 occurs then Q_2 also should also occur.

The frequent patterns or rules will be output if they are having support and confidence greater than or equal to the minimum support and confidence set up during the experimental analysis.

There are many methods to do association rule mining like apriori algorithm, partitioning method and FP growth method.

We have used the apriori algorithm in this paper. In apriori algorithm, we find the large frequency itemsets. In the first pass, we count the item occurrences to determine the large 1-itemsets. A subsequent pass, say pass k, consists of two phases. First, the large itemsets L_{k-1} found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the apriori-gen function. Next, the database is scanned and support of the candidates in C_k is counted. The apriori-gen function takes an argument L_{k-1}, the set of all large (k-1) itemsets. It returns a superset of the set of all large k-itemsets. It performs two functions. First in the join step, we join L_{k-1} with L_{k-1} .Next, in the prune step, we delete all itemsets c € C_k such that some (k-1) subset of c is not in L_{k-1}.

c. Predicted Queries

After mining the association rules, the queries that occur with the current query in the rules will be the output. This is because as per the rules the query in the output will have maximum probability to be fired by the user next to the current query. Also, the association rules go on updating periodically and the relation between the queries can be found to predict the next query.

V. EXPERIMENTAL EVALUATION

The proposed system has been implemented in Java. The following figures (Figure 3 to Figure 7) show the snapshots of the implementation.

1. Query Log is maintained for each user as shown in the Figure 3.

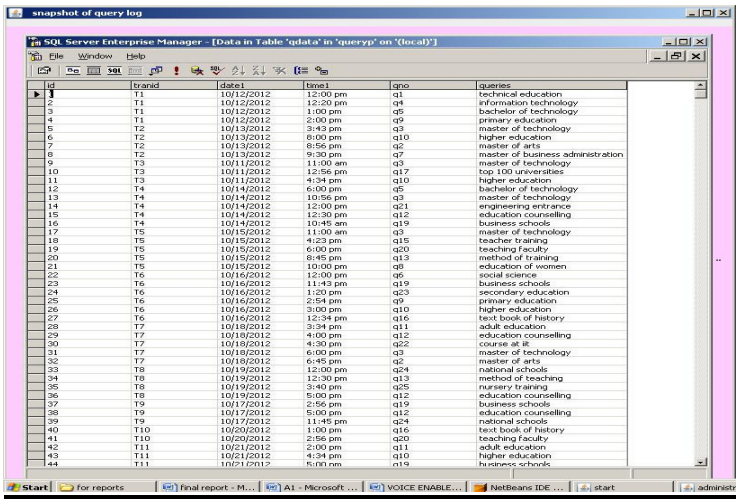

Figure 3: Snapshot showing how Query log of each user

2. Association rule Mining is applied on the query log maintained as shown and frequent querysets are found.

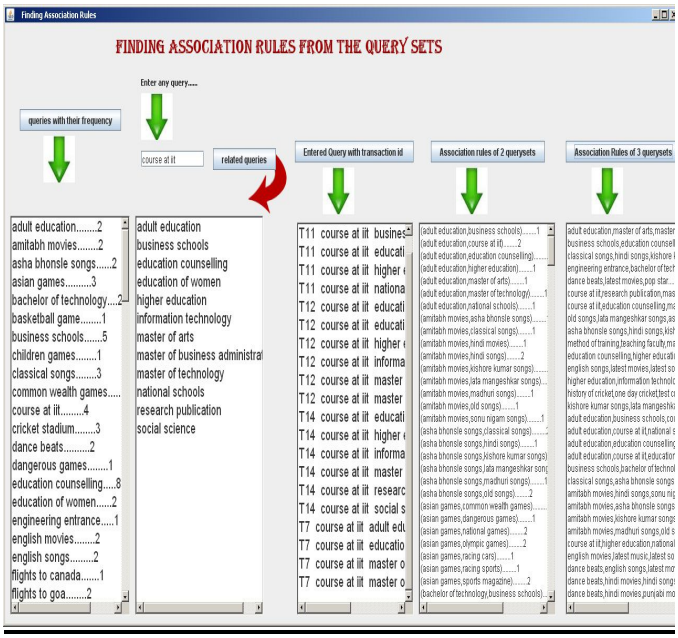

Figure 4: Snapshot showing how Association rules are mined and frequent query sets is found

3. After finding the frequent querysets, now the system finds the queries having maximum probability and which the user can fire next to his current query on the interface.

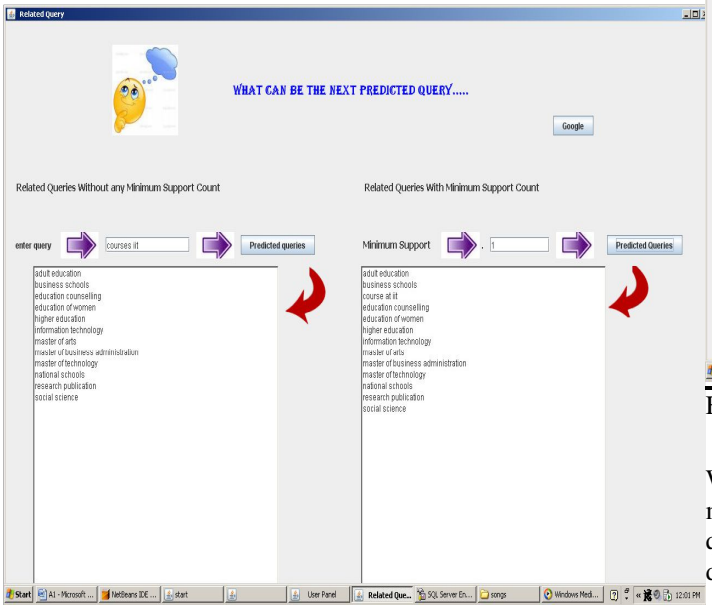

Figure 5. Snapshot showing how the next query is predicted

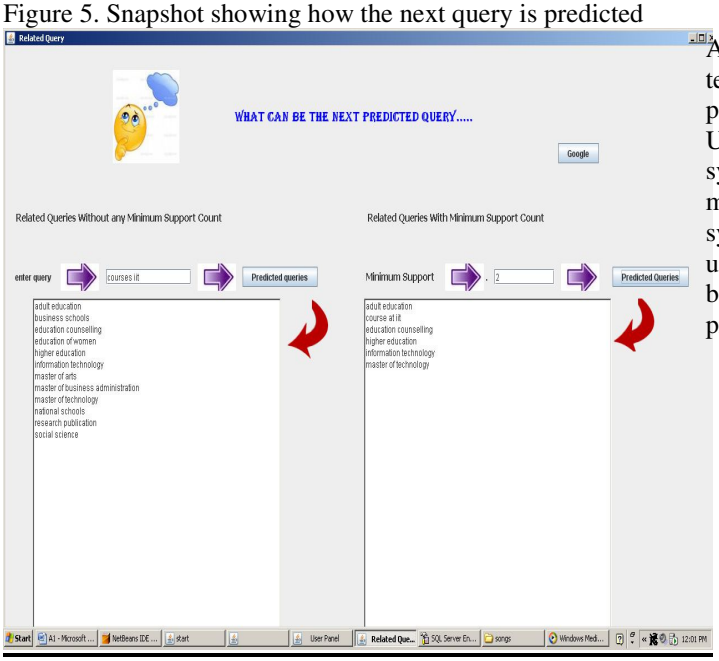

Figure 6. Snapshot showing how the next query is predicted

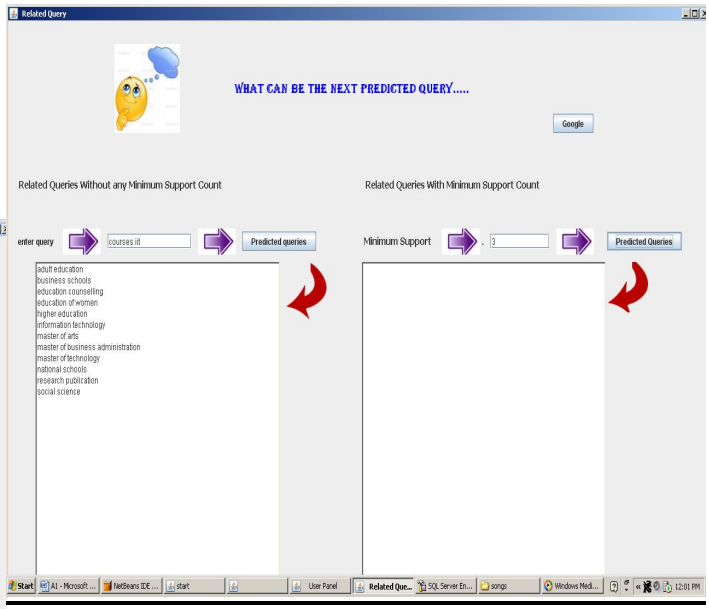

Figure 7. Snapshot showing how the next query is predicted

While implementation and experimental evaluation, a minimum threshold has been set for the support and all the queries that has the support value greater than or equal to the decided threshold value will be considered.

VI. CONCLUSION

Association rule discovery is one of the most important techniques in the field of data mining. It aims at finding patterns that exists in the databases.

Using the approach of Association rule mining, the proposed system works for the prediction of the query that the user is more likely to fire next to his current query. The proposed system worked well for some sample queries entered by the user and has shown promising results. The same approach can be used in search engine to improve its efficiency and thus providing better results to the user.

VII. REFERENCES

[1] Christopher D. , Prabhakar Raghavan and Hinrich Schütze "Introduction to information retrieval" , Cambridge Univeristy Press, 2008.

[2] Fonseca Bruno M. ,Golgher Paulo B., de Maura Edleno S., Ziviani Nivio, "Using association rules to discover search engine related queries", Proceedings of the First Latin American Web Congress (LA-WEB), IEEE, 2003.

[3] Gupta Deepti, Puniya Antima, Bhatia kumar komal, "Prediction of the Query of the Search Engine using Backpropogation Algorithm", IJCSE, 2011.

[4] Lin, K.H, "Predicting Next Search Actions with Search Engine Query Logs", Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference on (Volume: 1), 2011.

[5] Georges Dupret and Marcelo Mendoza, "Recommending Better Queries Based on Click-Through Data", LNCS, Springer, 2005.